

DOCUMENT RESUME

ED 261 075

TM 850 477

AUTHOR Kingsbury, G. Gage
TITLE A Comparison of Item Response Theory Procedures for Assessing Response Dimensionality.
PUB DATE Apr 85
NOTE 35p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1-4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Comparative Analysis; Elementary Education; Estimation (Mathematics); Factor Analysis; *Item Analysis; *Latent Trait Theory; *Mathematical Models; Multiple Choice Tests; Statistical Analysis; Statistical Studies
IDENTIFIERS Principal Axis Procedure; *Response Function Discrepancies; *Unidimensionality (Tests)

ABSTRACT

A procedure for assessing content-area and total-test dimensionality which uses response function discrepancies (RFD) was studied. Three different versions of the RFD procedure were compared to Bejar's principal axis content-area procedure and Indow and Samejima's exploratory factor analytic technique. The procedures were compared in terms of the consistency of decisions concerning dimensionality, and the amount of detail concerning content-area dimensionality which is available with each procedure. Achievement test data were from two multiple choice tests administered to over 2000 elementary school students in the Portland (Oregon) Public School system. One test contained 55 mathematics skill items and the other, 44 reading skill items. Results indicated that exploratory factor analysis provided a more global view of dimensionality; it was most likely to indicate that multidimensionality existed in the item responses, independent of the different content areas. The principal axis procedure tended to indicate more content-area multidimensionality. The RFD procedure was shown to be a viable method of determining response dimensionality, yet no single variation was consistently more informative. (GDC)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED261075

A COMPARISON OF ITEM RESPONSE THEORY PROCEDURES
FOR ASSESSING RESPONSE DIMENSIONALITY;

G. Gage Kingsbury
Portland Public Schools
Portland, Oregon

Paper presented at the annual meeting of the
National Council on Measurement in Education

Chicago, Illinois
April, 1985

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

G. G. Kingsbury

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to improve
reproduction quality.

* Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

TM 850 477

A COMPARISON OF ITEM RESPONSE THEORY PROCEDURES FOR
ASSESSING RESPONSE DIMENSIONALITY

G. Gage Kingsbury
Portland Public Schools

Most testing procedures based on item response theory (IRT: Lord, 1980) require that a unidimensional response model fit people's responses to the test questions. Techniques to assess the dimensionality of item responses are extremely important in choosing a response model, particularly when the items being calibrated are from achievement tests, which might well be multidimensional.

Lord & Novick (1968, p. 359-360) have shown that if the complete latent space is known, the item parameters and person parameters specified within this space must be invariant within a linear transformation. Estimates of these parameters may also differ due to the estimation procedures used and sampling fluctuations).

* This property leads to a set of procedures for determining whether test responses to an achievement test with multiple content areas are unidimensional. For any given item, parameter estimates may be obtained from the total test and from a set of items of more homogeneous content. If these estimates are the same (within fluctuation due to sampling and

Comparison of Procedures for Assessing Dimensionality

estimation) after the appropriate linear transformation, it can be taken as strong evidence that students responded to the items in different content areas using the same underlying trait. That is, the different content areas only tapped one underlying response dimension, and the dimension tapped was the same as that underlying performance on the total test.

Bejar, Weiss, & Kingsbury (1977) have suggested a regression procedure for determining whether parameter estimates obtained within content areas deviate more than expected from parameters estimated from the total test. Bejar (1980) modified this procedure to allow the detection of differential dimensionality among content areas on a test by comparing discrepancies of the parameter estimates from a theoretical axis and a principal axis fit directly to the data in each content area. This procedure allows the researcher to obtain evidence concerning which areas of content tap different underlying dimensions, if any do.

These procedures approach dimensionality in a direct manner since they address the issue of content specificity. There are at least two areas in which these procedures need to be improved, however. First, these procedures address each item parameter singly, while the important consideration is whether the response function as a whole differs from total-test to content-area calibration. Second, content-area calibrations are based on fewer items than the total-test calibrations. This may cause differential accuracy in the item calibration (Swaminathan & Gifford, 1980) which may distort the analysis of dimensionality, and particularly the analysis of differential content-area dimensionality proposed by Bejar (1980).

Comparison of Procedures for Assessing Dimensionality

This paper will describe a procedure for the assessment of content-area and total-test dimensionality which will attempt to address these two areas through the use of response function discrepancies. In addition, this paper will compare this proposed procedure to two procedures suggested in the literature to determine its operating characteristics.

RESPONSE FUNCTION DISCREPANCIES: A METHOD TO ASSESS DIMENSIONALITY

If an item is calibrated assuming a probabilistic item response model, a response function is derived which functionally relates an examinee's trait level to the probability that s/he will answer the question correctly. If the item is calibrated twice, once within the total test and once within a selected subset of homogeneous items, the calibrations may differ. In turn, this will cause the probabilities of observing a correct response to the item to differ, depending on the context in which it is encountered. This violation of the assumption of local independence (Lord & Novick, 1968) would serve as evidence of a multidimensional trait space. This type of dimensionality analysis has been used by Bejar, Weiss, & Kingsbury (1977) and by Bejar (1980). This paper describes a somewhat different approach to the same technique.

In order to evaluate the complete response function, the absolute difference between expected probabilities of a correct response to an item estimated in total-test and content-area calibrations may be accumulated across the effective range of

Comparison of Procedures for Assessing Dimensionality

the underlying continuum to determine the area between the two response functions (via integration or a discrete approximation). This response function discrepancy (RFD) summary statistic for each item would serve to describe the impact of different parameter estimates on the expected response probabilities, if such differences occurred.

In order to correct for the effect of differential item sample sizes on parameter estimation, it is necessary to establish a reference calibration which has the same number of items. Thus, if 20 items from a 50-item test are to be included in a particular content area calibration, then 20 items should be chosen representatively from the total test to serve as a reference calibration. The reference-test parameters serve as an expected calibration differing from the total-test calibration only due to the reduction in the number of items.

Combining these two concepts, a new procedure for the assessment of dimensionality with content-area response functions may be described. Denote the area between total-test and content-area response functions as RFD_{TCi} (for each item i) with mean $MRFD_{TC}$, and the area between the total-test and reference-test response functions as RFD_{TRi} with mean $MRFD_{TR}$. One may examine dimensionality within a content area by performing an independent-groups t-test between $MRFD_{TC}$ and $MRFD_{TR}$ from the appropriate length reference test. For more detailed analysis of dimensionality, each value of RFD_{TCi} could be compared to $MRFD_{TR}$, allowing for appropriate standard errors.

Comparison of Procedures for Assessing Dimensionality

The RFD procedure should serve as an adjunct to the procedures described by Bejar, Weiss and Kingsbury (1977) and Bejar (1980) to determine whether a set of test responses is unidimensional via content area assessment. The new procedure provides an overall index for each item and adds an adjustment for the calibration error to be expected with small item sets.

A COMPARISON OF RFD PROCEDURES WITH OTHER TECHNIQUES

In order to determine the working characteristics of the RFD procedure, this study compared several versions of the RFD procedure to the content-area procedure suggested by Bejar (1980), and to an exploratory factor analytic technique for assessing dimensionality (suggested by Indow & Samejima, 1966 and Lord & Novick, 1968). Procedures were compared in terms of the consistency of decisions concerning dimensionality, and in terms of the amount of detail concerning content-area dimensionality available with each procedure.

Method

Subjects and tests. Data used in this study came from the administration of two achievement level tests to samples of over 2000 grade school students in the Portland Public School system. The two tests used were multiple-choice tests, one with 55 mathematics skill items and the other with 44 reading skill items, respectively. Each of these tests served as a survey test for the students in the appropriate grade level, and so consisted of items which tapped a number of discrete instructional goals. All items which were designed to tap a given goal were considered to represent a single content area for purposes of this study (although a finer breakdown of items

Comparison of Procedures for Assessing Dimensionality

into subgoals is also possible).

For the reading test four content areas were defined. These were 1)word meaning (11 items), 2)literal comprehension (14 items), 3)interpretive comprehension (11 items), and 4)evaluative comprehension (8 items). For the math test six content areas were defined which were 1)multiplying whole numbers (5 items), 2)dividing whole numbers (5 items), 3)numeration (10 items), 4)computations with fractions and decimals (10 items), 5)geometry, measurement, and graphs (15 items), and 6)solving problems (10 items).

Since multiple-choice questions were being evaluated, an appropriate unidimensional IRT model to use was the three-parameter logistic model (Lord & Novick, 1968). Maximum likelihood estimates of the parameters of this model were used throughout the remainder of this study.

Procedures evaluated. Five procedures for the assessment of unidimensionality were used and compared with each of the tests. Three of these procedures were variants of the RFD procedure described above. These three procedures were designed to determine how sensitive the RFD procedure is to the range of the achievement (theta) metric used in its calculation. The first of these (RFD1) evaluated the differences between estimated item response functions by using a discrete approximation across the interval from -1.0 to +1.0 along the theta metric. The second procedure (RFD2) used the same type of discrete approximation evaluated across the interval from -2.0 to +2.0 along the theta metric. The third of these procedures (RFD3) evaluated the discrete approximation from -3.0 to +3.0 along the theta metric.

Comparison of Procedures for Assessing Dimensionality

It was anticipated that as the interval used in the RFD calculation increased, differences in the lower-assymtote (c) parameter estimate would begin to dominate the RFD value. As an example, consider an item with exactly the same discrimination (a) and difficulty (b) parameter values but slightly different c estimates in total-test and content-area calibrations. If the RFD were evaluated over a small range of the achievement continuum, a very small value would be obtained, while if a very large range were used, the RFD would take on a much larger value. To determine the extent of this problem, the three ranges for the RFD calculation were chosen.

The fourth procedure used was a principle axis (PA) procedure described by Bejar (1980). This procedure involved computing the principle axis from a plot of content-area versus total-test parameter estimates, and determining how discrepant the principle axis was from the theoretical axis in each content area. In addition, this procedure included evaluating the mean-squared distance of each item plotted from the principle and theoretical axes.

The fifth procedure used was an exploratory factory analysis (EFA) of the type suggested by Indow and Samejima (1966) and modified by McBride and Weiss (1974). This procedure involved a principle axis factor analysis of a matrix of tetrachoric correlations obtained from responses to the entire test. Eigenvalues for factors extracted were plotted along with eigenvalues from a parallel analysis of random data. These eigenvalues were examined to determine the dimensionality of the response space.

Comparison of Procedures for Assessing Dimensionality

Analysis. For both tests, the five procedures were applied and compared in terms of decisions regarding overall dimensionality and the clarity of the decision. In addition, the three RFD procedures and the PA procedure were compared in terms of content-area specific decisions concerning dimensionality. The procedures were compared in terms of the consistency of decisions and the amount of information provided regarding the content-areas.

RESULTS

Exploratory Factor Analysis (EFA)

The factor analysis of responses to the reading test resulted in one factor with an eigenvalue greater than one (h^2 equal to 3.73). The mean communality with the extraction of ten factors was .157.

For responses to the math test, two factors were extracted with eigenvalues greater than one (h^2 equal to 4.40 and 1.77). Extracting ten factors, the mean communality was .224.

Figures 1 and 2 show the eigenvalues of the first ten factors extracted from responses to the reading and math tests, respectively. In addition, the eigenvalues for factors from comparable random data sets are also shown. These figures indicate somewhat different factor structures for the two tests.

From Figure 1 it can be seen that only the first factor extracted from the reading test has an eigenvalue that differs noticeably from that for the corresponding random-data factor.

Figure 1

Eigenvalues for reading and random factors

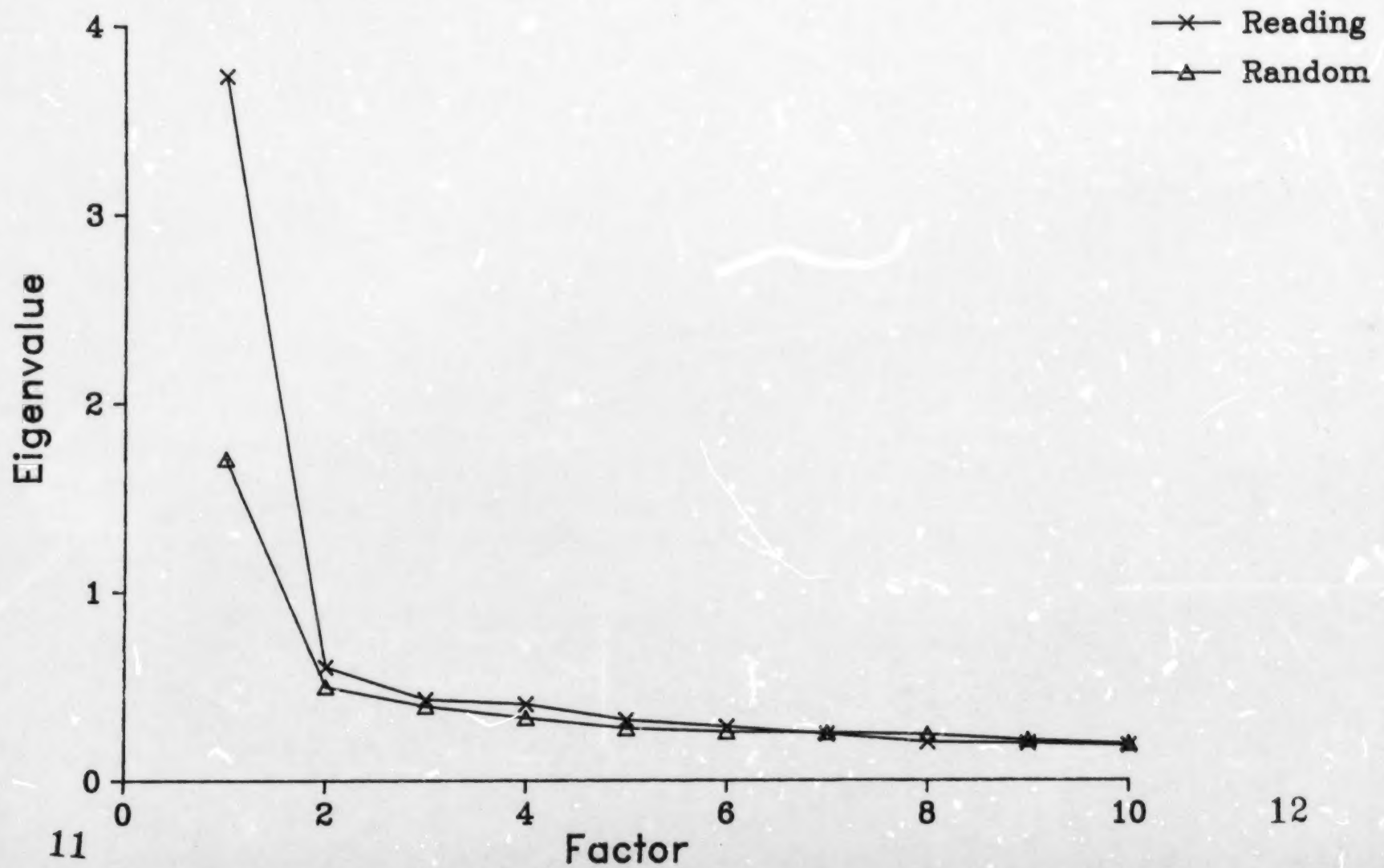
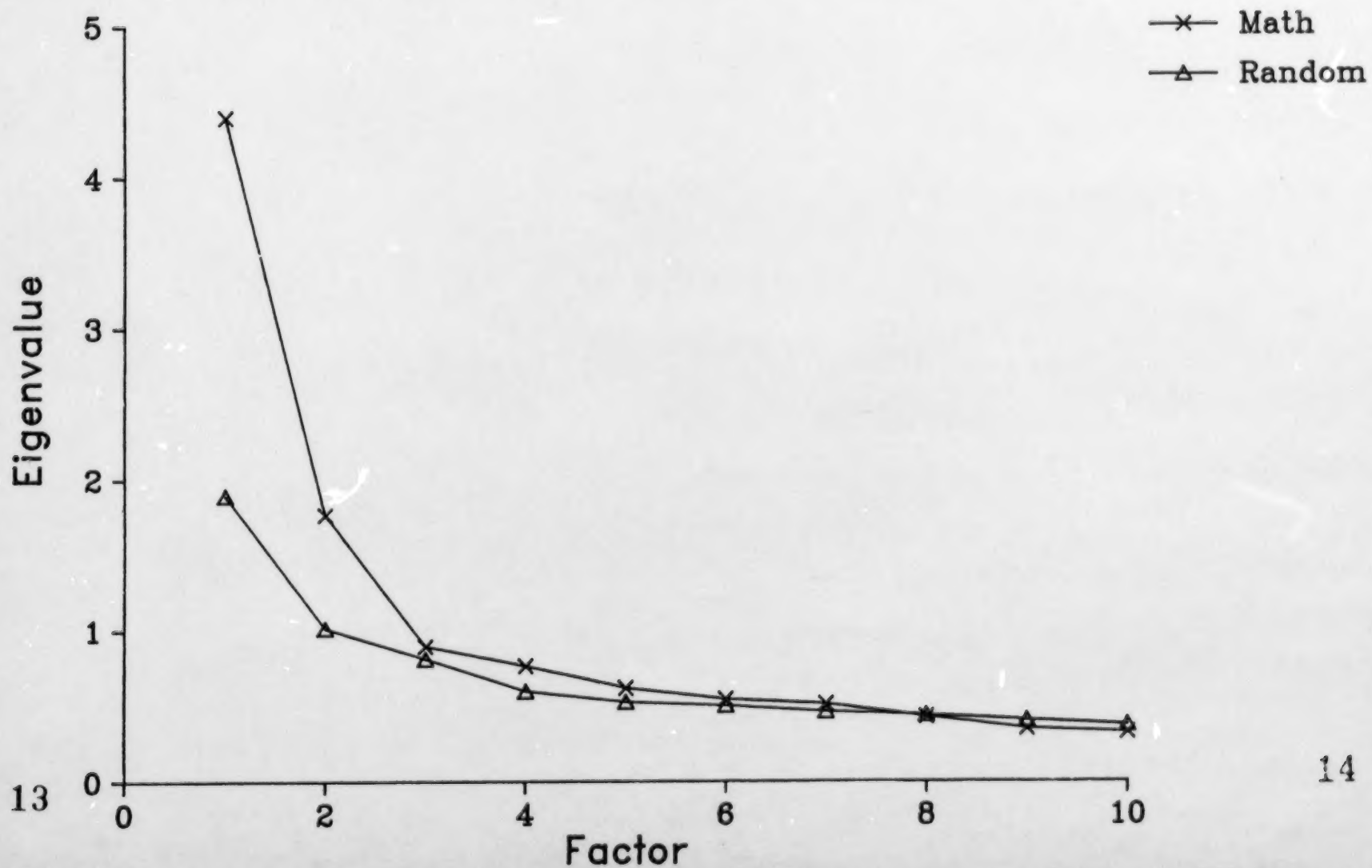


Figure 2

Eigenvalues for math and random factors



Comparison of Procedures for Assessing Dimensionality

For the second through tenth factors, the eigenvalues for real-data and random-data factors are virtually indistinguishable. According to the McBride & Weiss (1974) criteria, this pattern indicates that at most one important factor is present in the reading data.

For the math test, the first two real-data factors had eigenvalues noticeably higher than those obtained from the first two random-data factors. This result indicates that at most two important factors are present in the math test response data.

An additional point should be noted concerning these factor analytic results. The low communalities in the responses to items on both tests indicate that most of the item score variance was unique to the individual items, and not due to shared or common dimensions. The factor structures in these data sets were not very strong, which might have effected item parameter estimation.

In summary, the factor analytic assessment of dimensionality indicated that responses to the reading test were mitigated by a single, weak common factor while responses to the math test were mitigated by two weak common factors. Thus, a unidimensional IRT model might not fit the math test responses, while it might fit the reading test responses somewhat better.

Principal Axis (PA) Analysis

Comparison of Procedures for Assessing Dimensionality

Figures 3 and 4 show the total-test and content-area estimates of the b parameter obtained for each item on the reading and math tests, respectively. In addition, the principal and theoretical axes are also shown in these figures. Each of these figures shows a clear, strong relationship between b -parameter estimates obtained within the total test and those obtained within the individual content areas. For reading, the Pearson product-moment correlation between the parameter estimates was .966, while for math it was slightly lower at .901. The parameter estimates from the math test items tend to be more divergent in total-test and content-area calibrations than those for the reading test items. This could indicate a greater tendency toward multidimensionality in the math test responses.

It can also be seen from these figures that the principal and theoretical axes are quite similar. For each test, the principal axis lies just above the theoretical axis with a nearly consistent difference at all observed difficulty levels. The slope of the principal axis for the reading test was 1.004 with an intercept of .129. For the math test, the principal axis had a slope of .980 and an intercept of .213. These can be compared to a slope of one and an intercept of zero for the theoretical axis.

Figures 5 and 6 show the mean squared distance for items in each content area from its principal and theoretical axes (Note that a separate principal axis was calculated for each content area for this comparison.) Large differences between distances to the principal and theoretical axes in a content area would indicate that items in the content area might be linked by a "content-area" factor, implying that some

Figure 3

Reading total test and content area b-estimates

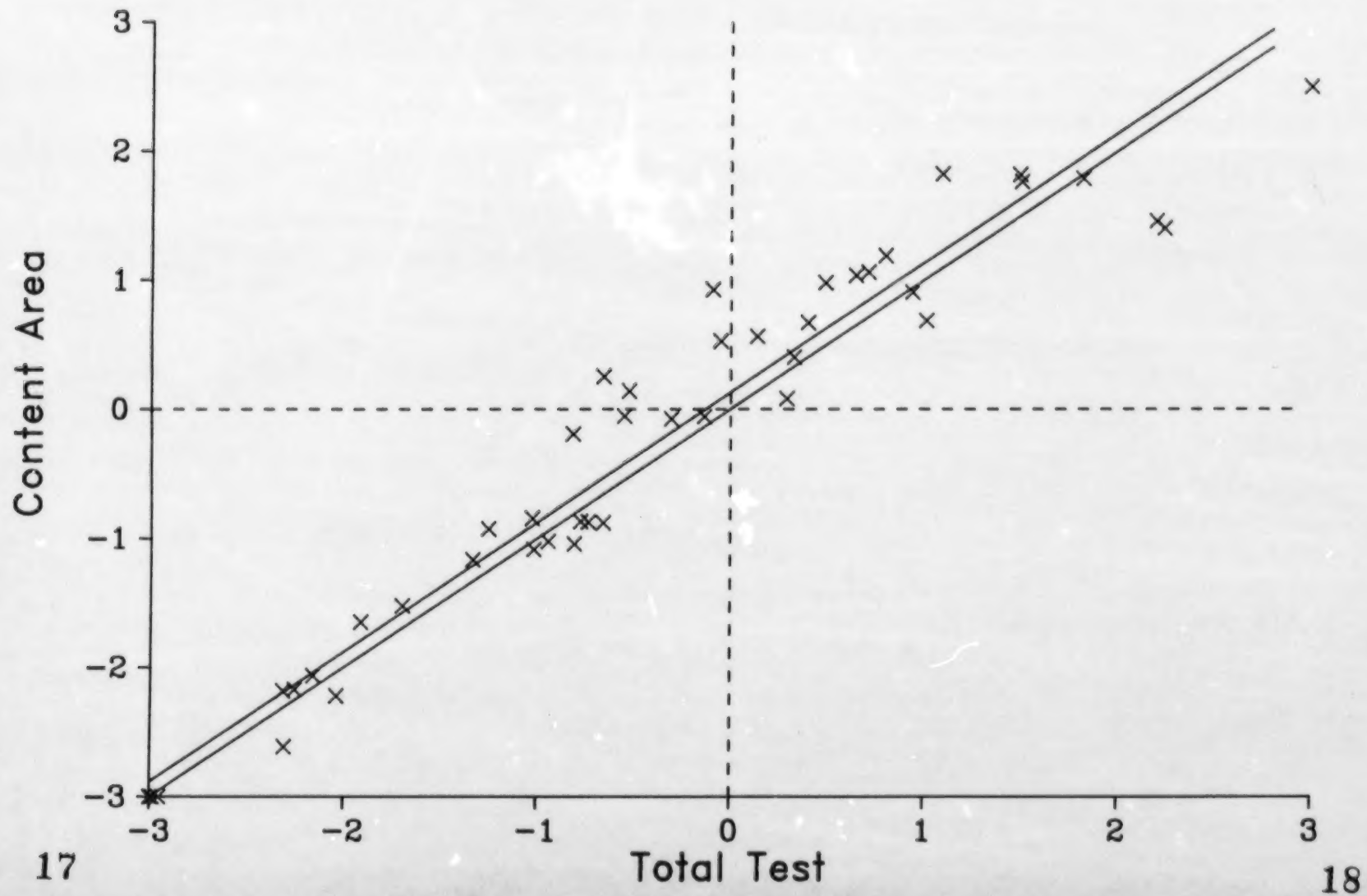


Figure 4

Math total test and content area b-estimates

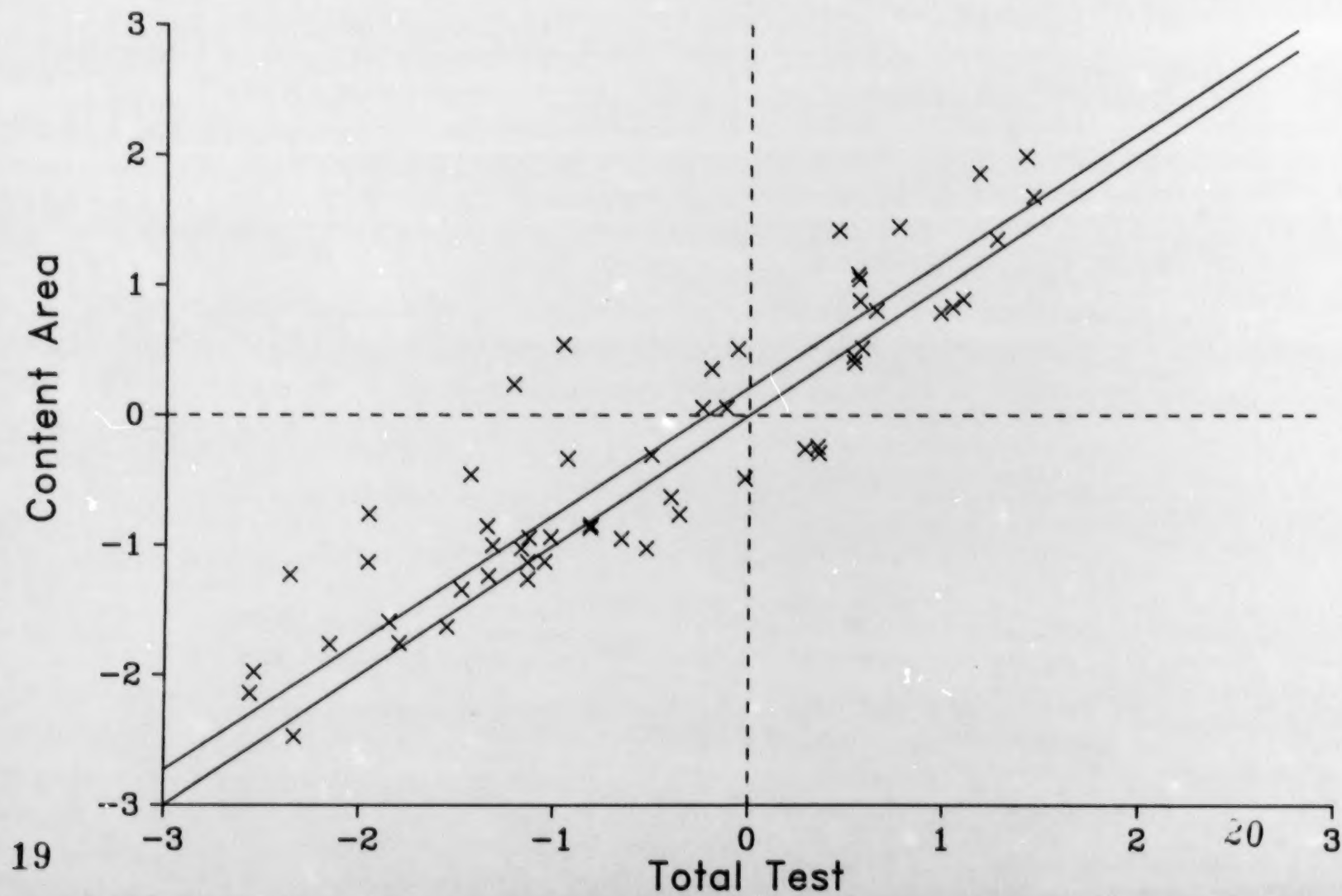


Figure 5

Mean squared distances to axes for reading

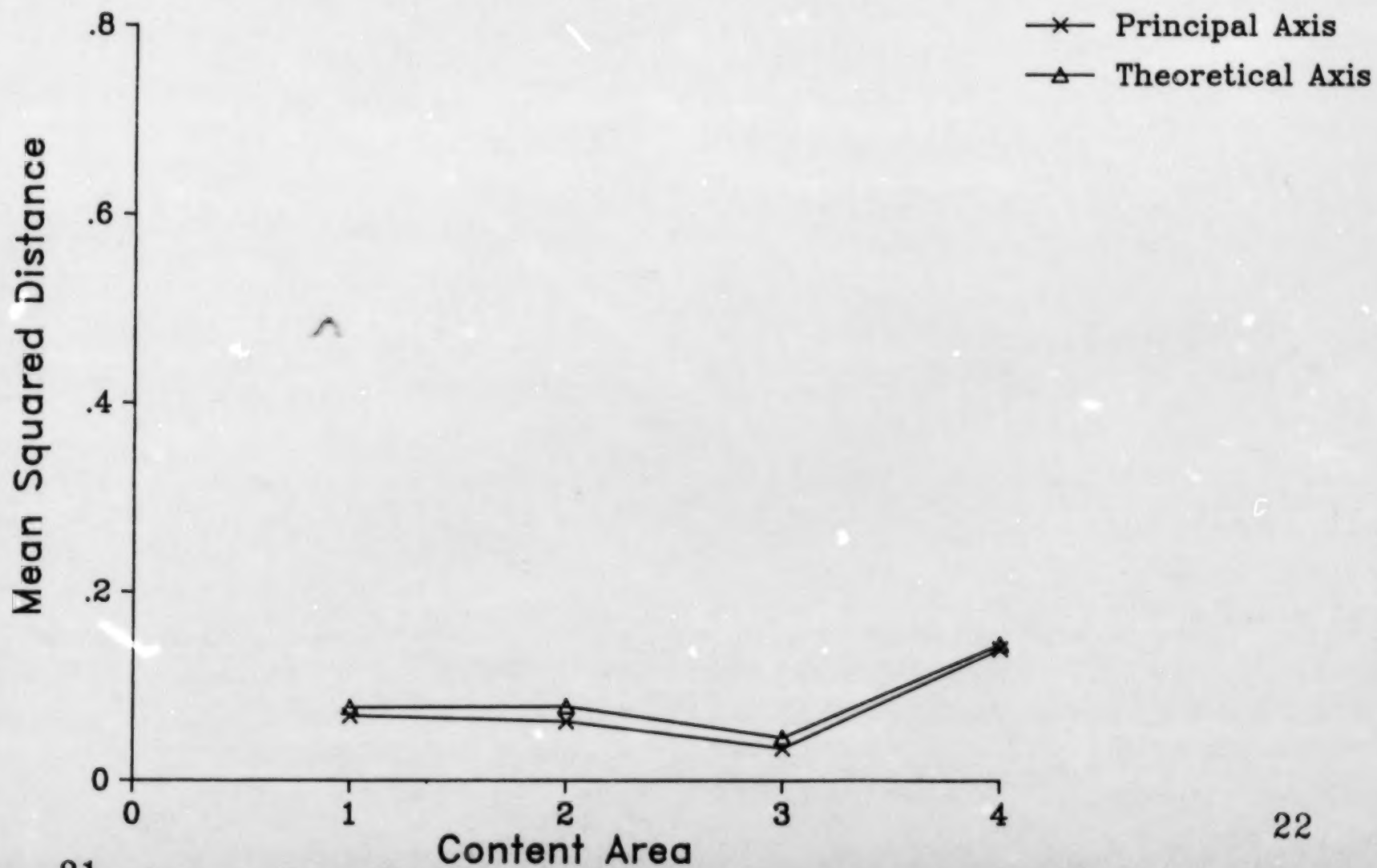
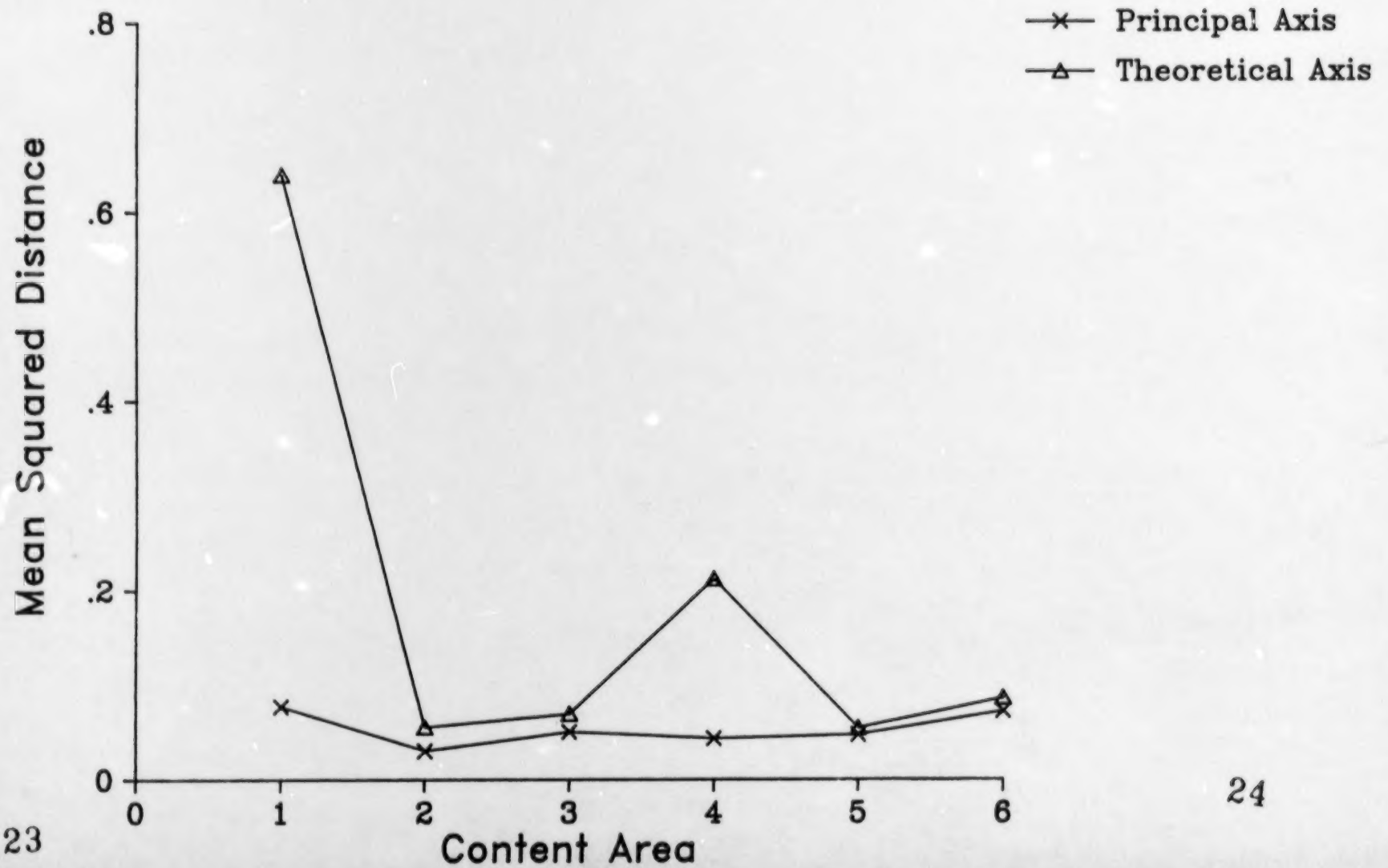


Figure 6

Mean squared distances to axes for math



Comparison of Procedures for Assessing Dimensionality

multidimensionality exists in the data.

Figure 5 shows that, for the reading test, the distances to the principal and theoretical axes were very similar in all content areas. Content area 4 resulted in the highest average squared distance to both axes. These results show very little evidence of multidimensionality, although content area 4 did exhibit some divergence of the b-parameter estimates.

Figure 6 shows a different pattern of mean squared distances for the math test. For content areas 1 and 4, the mean squared distances to the theoretical axis were noticeably greater than those to the principal axis. For these two content areas, the differences in the mean squared distances to theoretical and principal axes are greater than the mean squared distance to any axis in any other content area.

In summary then, the principal axis analysis indicated very small differences in slope between the theoretical axis and the principal axis across content areas. When individual content areas were considered, the reading test showed no evidence of multidimensionality other than a larger dispersion of the b-parameter estimates in content area 4. The math test showed some evidence of multidimensionality in content areas 1 and 4.

Response Function Discrepancy (RFD) Analysis

Using the two tests containing a total of ten different content areas and the three variations of response function discrepancy, a total of thirty independent-groups t-tests were done to compare the mean RFD observed in each content area

Comparison of Procedures
for Assessing Dimensionality

($MRFD_{TC}$) to the mean RFD observed in the corresponding representative item sample ($MRFD_{TR}$). In the thirty cases considered, the largest difference observed between the content-area and representative sample means had a p-value greater than .05. In addition, 29 of the 30 t-tests resulted in p-values greater than .25. Thus, the RFD analysis indicated no evidence of multidimensionality within any of the content areas within either test with RFD1, RFD2, or RFD3.

It is important to note that while none of the differences in mean RFD observed between $MRFD_{TC}$ and $MRFD_{TR}$ were significantly different from zero, 26 of the thirty values of $MRFD_{TC}$ were significantly different from zero (with $p = .01$) when considered alone. Therefore, the apparent evidence of multidimensionality (significant values of $MRFD_{TC}$) was removed through the use of an appropriate reference test of the same length.

Since no evidence of multidimensionality was found through this initial phase of the RFD analysis, the procedure would normally be terminated at this point with a finding of no apparent evidence of multidimensionality stemming from the different content areas on the test. However, since this is an explanatory study, and since one goal of the study is to discover the operating characteristics of the RFD variants, it is informative to continue the analysis.

Figures 7 and 8 show the mean values of RFD1, RFD2, and RFD3 for each content area on each test. The three RFD variants have been scaled by the variance of the corresponding reference test in order to allow comparisons on the same scale.

Figure 7

Mean absolute RFD for each reading content area

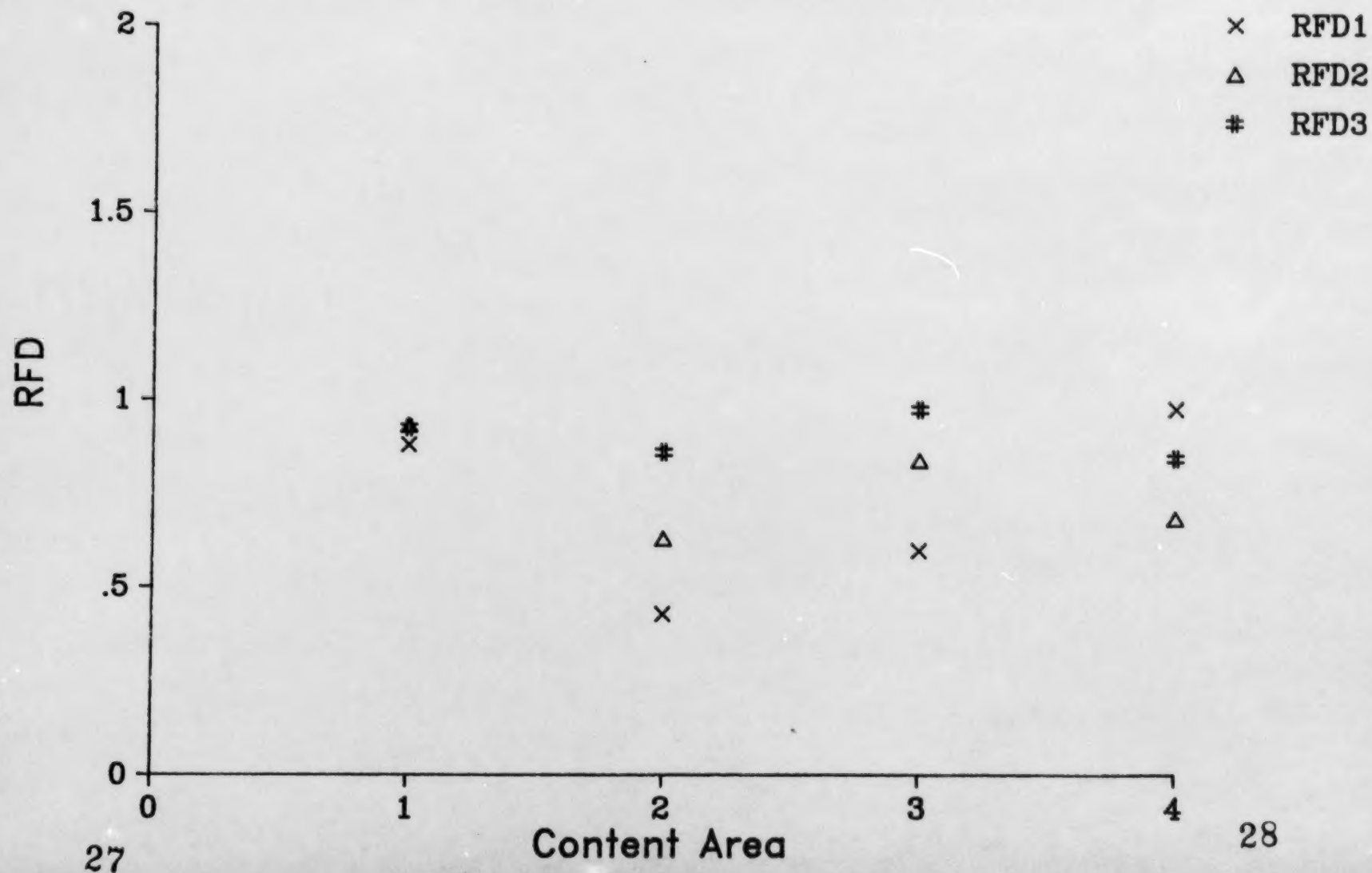
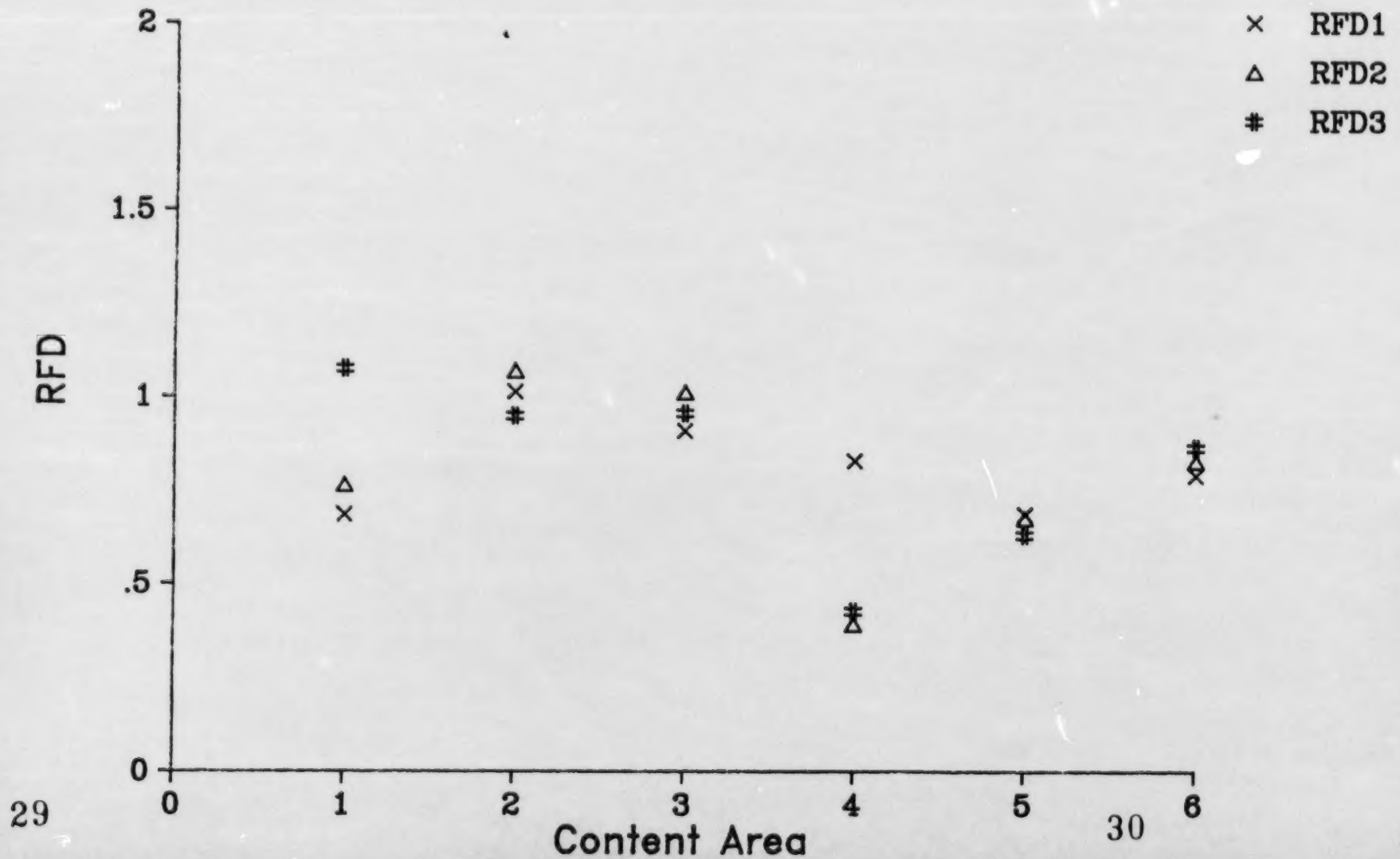


Figure 8

Mean absolute RFD for each math content area



Comparison of Procedures for Assessing Dimensionality

For the reading test, Figure 7 shows some small differences among content areas. In general, the differences between RFD variations shown in this figure tend to be much larger than the differences between content areas. RFD1 yields the most discrepant results between content areas, with content areas 1 and 4 having the largest values, and therefore the most evidence of multidimensionality. RFD2 yields less variation between content areas, although the level of response function discrepancy observed is, in general, larger than that for RFD1. RFD3 tends to have the largest RFDs, with the least difference between content areas.

Figure 8 shows a somewhat different pattern for the math test. The first trend to be noted was that, except for content areas 1 and 4, differences between the RFD variants were very small, and were much smaller than differences between content areas. For content area 1, RFD3 results in a higher value than the other two variants. On the other hand, for content area 4, RFD1 results in the highest scaled response function discrepancy. It is interesting to note that these odd content areas are the same areas that the principal axis procedure picked as having some evidence of multidimensionality. While no evidence of multidimensionality is found for these content areas using the RFD procedure, they do seem to have odd characteristics. To show the contrast between these two procedures, content area 4 was selected by the principal axis procedure as exhibiting some evidence of multidimensionality, while it had some of the lowest RFD values observed, thus having the least evidence of multidimensionality.

Comparison of Procedures for Assessing Dimensionality

In summary, the t-tests conducted as the opening procedure of the RFD analysis showed no evidence of content-area based multidimensionality. No clear indication of a "best" RFD variant was discovered, because the patterns differed between tests a great deal. This may be due in part to the dependence of the RFD values on the difficulty of the items.

CONCLUSIONS AND DISCUSSION

Results of this study concern two major topics: 1) comparisons of the RFD procedure to the EFA and PA procedures and, 2) comparisons of the variations of the RFD procedure.

The comparisons of RFD and EFA are not clear-cut because the procedures result in different types of information. RFD procedures provide content-area information which the EFA procedure does not, but the EFA procedure provides a more global view of dimensionality. To the extent that multidimensionality exists in the item responses independent of the different content areas, the EFA procedure might reveal it while the RFD (and PA) procedure might not.

This study may have shown an example of this phenomenon. The EFA procedure indicated that the math test responses might be mitigated by two common factors. No indication of this multidimensionality was found with the RFD analysis. The second math factor may be a factor unrelated to content area (and in fact, an examination of the salient loadings on this second factor do not indicate any relation to a particular content area). This type of factor would be detected by the

Comparison of Procedures for Assessing Dimensionality

EFA, but not by the content area procedures.

Finally, comparing the RFD variants to the PA procedure, it was shown that the two procedures reached somewhat different conclusions, with the PA procedure tending to indicate more content-area multidimensionality. A contributing factor to the differences might be the use of the reference-test procedure with RFD. The RFD procedure may tease out differential parameter estimation accuracy that the PA procedure leaves in when determining dimensionality. This hypothesis is supported by the finding that most of the RFD values were significantly different from zero, but none were significantly different from the corresponding RFDs from the reference calibrations.

The RFD procedure has been shown to be a viable method of determining response dimensionality but no single variation was shown to be consistently more informative across tests. When used in conjunction with other procedures such as EFA, the RFD procedure may serve to provide a more detailed, complete picture of the response space evoked by achievement tests. Further research into the RFD procedure and its variants seems warranted. One promising variation that might be useful is the calculation of the RFD relative to the total-test b-parameter estimate of each individual item, rather than relative to a given interval on the trait continuum.

It is apparent that further research concerning the variants of the RFD procedure might be best done in simulation studies which would allow the interpretation of small differences in RFD values against an absolute standard. In contrast, further comparisons of the techniques for assessing dimensionality might be done either in simulation or in

Comparison of Procedures
for Assessing Dimensionality

live-testing situations. The simulations would again allow the use of an absolute standard, but the live-testing studies would allow one to investigate how well the procedures perform with actual conditions.

REFERENCES

- Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.
- Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. (1977). Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Indow, T. & Samejima, F. (1966). On the results obtained by the absolute scaling model and the Lord model in the field of intelligence. Yokohama: Psychological Laboratory, Hiyoshi Campus, Keio University.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum.
- Lord, F. M. & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.
- McBride, J. R. & Weiss, D. J. (1974). A word knowledge item pool for adaptive ability measurement (Research report 74-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Swaminathan, H. & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.) New horizons in testing. Latent trait test theory and computerized adaptive testing. New York: Academic Press.